Proceedings of the 2017 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI 2017), 16-18 November 2017, Daegu, South Korea
DOI: 10.1109/MFI.2017.8170443

Data Analytics Development of FDR (Flight Data Recorder) Data for Airline Maintenance Operations

Chang-Hun Lee, Hyo-Sang Shin, Antonios Tsourdos, and Zakwan Skaf

***Abstract*— In this article, we propose a data analytics development to detect unusual patterns of flights from a vast amounts of FDR (flight data recorder) data for supporting airline maintenance operations. A fundamental rationale behind this development is that if there are potential issues on mechanical parts of an aircraft during a flight, evidences for these issues are most likely included in the FDR data. Therefore, the data analysis of FDR data enables us to detect the potential issues in the aircraft before they occur. To this end, in a data pre-processing step, a data filtering, a data sampling, and a data transformation are sequentially performed. And then, in this analysis, all time series data in the FDR are classified into three types: a continuous signal, a discrete signal, and a warning signal. For each type of signal, a high-dimensional vector by arranging the time series data is chosen as features. In the feature section process, a correlation analysis, a correlation relaxation, and a dimension reduction are sequentially conducted. Finally, a type of k-nearest neighbor approach is applied to automatically identify the FDR data in which the unusual flight patterns are recorded from a large amount of FDR data. The proposed method is tested with using a realistic FDR data from the NASA's open database.**

I. INTRODUCTION

Recently, most of aircraft have digital FDRs (flight data recorders) [1] in order to record phenomena that occur on an aircraft during a flight. In the FDR data, there are sensor measurements of some parts of the aircraft with a purpose of health monitoring, such as an engine fan speed, an engine vibration, an oil temperature, an acceleration, and so on. Additionally, the FDR data include a flight environment data and a flight status, such as Mach number, an altitude, a pressure, a speed, and so on. In airline companies, the FDR data is important in terms of aircraft maintenance and aircraft operation. If there are technical issues on certain parts of an aircraft during a flight, evidences of these issues are most likely recorded in the FDR data. If we properly analyze the FDR data, then the analysis results of FDR data allow us to detect the potential issues inside the aircraft before they occur. Hence it can be utilized for a high-level fault diagnosis so that it can help improve the airline maintenance operations.

Accordingly, most airline companies have operated a dedicated department intended to analyze and utilize the FDR data. However, because of a huge amount of FDR data, it is difficult for a human operator to consume all the FDR data within a limited time. Therefore, recently, there have been research activities to develop an automated data analysis algorithm for the FDR data using data mining techniques and computing resources.

In a reference [2], a data analysis algorithm was proposed to detect atypical flights in sequences of discrete flight parameters. In this paper, the discrete parameters in the FDR data were only considered. As a remedy, anomaly detection in the FDR data was suggested by using the one-class SVM (support vector machine) with the multiple kernel [3]. In this paper, heterogeneous parameters, continuous and discrete, were handled together by discretizing the continuous parameters. This algorithm is a type of supervised learning methods so that it requires some training data. In a reference [4], a clustering-based algorithm, which is one of unsupervised learning methods, was proposed to detect an abnormal operation of an aircraft using the FDR data.

In this paper, we propose a new data analysis algorithm for the FDR data with a different aspect compared to the existing works. This work is motivated by the previous work [4]. Unlike the previous work, this study focuses on using the FDR data from the perspective of improving the airline maintenance operation, not the aircraft operations. In the reference [4], the Euclidian distance was used for a similarity measure of the continuous parameters and the discrete parameters, regardless of the parameter characteristics. Therefore, the similarity measure of discrete parameters may not be correctly computed in this approach. Also, in [4], a density-based clustering algorithm called DBSCAN (density-based spatial clustering of applications with noise) [5] was applied, but an output of this algorithm is sensitive to the design parameters regarding a density criteria of this algorithm. As a remedy, in this paper, time series data in the FDR are classified into three categories: a continuous, a discrete, and a warning signal, according to their parameter characteristics. A type of k-nearest neighbour approach is applied to detect the FDR data in which unusual flight patterns are recorded.

In a data pre-processing, a data filtering is performed using a moving median filter [6] and a second-order linear filter [6] in order to reject outliers and noises in the time series data. A data sampling is performed in order to make each FDR data to be mapped into a comparable data space. A data transformation is carried out to uniformalize value ranges of parameters. In a feature generation process, a feature is chosen as a high-dimensional vector by arranging the time series data for each type of signals. In order to generate this feature, we first perform the Pearson correlation analysis [6]

Chang-Hun Lee is with the School of Aerospace, Transport and Manufacturing, Cranfield University, Cranfield, MK43 0AL (corresponding author to provide e-mail: lckdgns@gmail.com).

Hyo-Sang Shin is with the School of Aerospace, Transport and Manufacturing, Cranfield University, Cranfield, MK43 0AL.

Antonios Tsourdos is with the School of Aerospace, Transport and Manufacturing, Cranfield University, Cranfield, MK43 0AL.

Zakwan Skaf is with the Integrated Vehicle Health Monitoring (IVHM) Centre, Cranfield University, Cranfield, MK43 0AL.

between all the continuous parameters to check redundant information in the FDR data. Base on a correlation analysis, a correlation relaxation is performed to weaken a multi-collinearly. The PCA (principal component analysis) [6] is performed to reduce a dimensionality of feature vector. In order to show performances of proposed method, it is tested with a realistic FDR data which is obtained from the NASA's open database.

This paper is composed as follows. In Section II, an overall process of proposed method is explained. In Section III, the proposed method is provided. In Section IV, the proposed method is tested with the FDR data. Finally, conclusions of our study are given in Section V.

II. The Overview of Proposed Data Analytics

A. Concept of Proposed Data Analytics

In the FDR data, each parameter is represented by the time series data. In this study, the time series data are classified into the three parameter types, according to their characteristics. For each parameter type, the time series data are transformed into the high-dimensional vector by arranging the time series data. Through this step, patterns of time series can be mapped into vectors with numerical values. The high-dimensional vectors of each FDR data are then compared each other. In this development, we assume that a majority of high-dimensional vectors exhibit a common pattern and a minority of high-dimensional vectors is far from the common pattern. By applying a type of k-nearest neighbor approach, the proximate vectors that can get together and the standalone vectors in a hyperplane are identified. Here, for each parameter type, different similarity measures are applied according to their characteristics. Finally, in this concept, such standalone vectors are regarded as the candidates of unusual patterns.

B. Overall Process of Proposed Data Analytics

This section explains the overall process of proposed data analytics shown in Fig. 1. The raw FDR data is composed of a set of the time series data. In the first step, the raw data needs to be pre-processed as

- **Data Filtering**: to reject outliers and remove noises in the time series data.
- **Data Sampling**: to match up a sampling rate of each parameter and to make different the FDRs become comparable.
- **Data Transformation**: to uniformalize value ranges of each parameter.

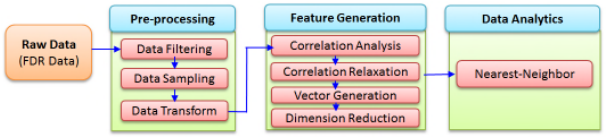


Figure 1. The overall process of proposed data analytics for the FDR data.

In this study, the feature is chosen as the high-dimensional time series vector formed by arranging all the sampled parameters. In order to produce this feature, the following sub-processes are performed.

- **Correlation Analysis**: to check redundant information; identify sets of parameters that are highly correlated each other.
- **Correlation Relaxation**: to weaken the effect of multi-collinearity among the highly correlated parameters.
- **Vector Generation**: to form the high-dimensional vector by arranging all the sampled parameters.
- **Dimension Reduction**: to reduce the dimension of high- dimensional vector using the PCA [6].

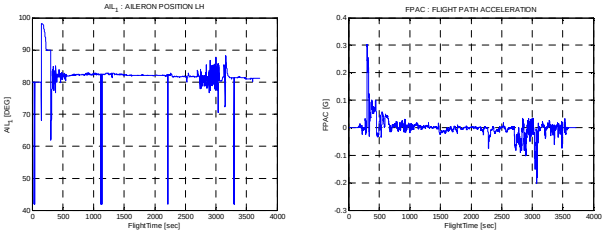
In the proposed data analysis process, the anomaly detection algorithm based on the nearest neighbor is used. The outputs of this data analytics are the information on a list of the FDR data that exhibits the unusual flight patterns to the others as well as the information on which parameters cause these unusual patterns. This information enables a high-level fault diagnosis for the airline maintenance operations.

III. Proposed Data Analytics Development

A. Understanding of Data

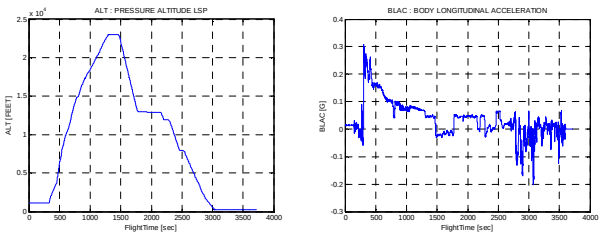
The FDR data used in this study are obtained from the NASA's open database [7]. Each FDR data consist of 186 different parameters including the continuous signals, the discrete signals, and the warning signals, respectively. Each parameter contains the information on a parameter name, a parameter type, a parameter sampling rate, a parameter unit, and a parameter description, respectively. In each FDR data, there are the three parameter types: the continuous, the discrete, and the warning signals. Also, the discrete parameters can be further categorized into an ordinal parameter and a binary parameter, respectively.

One of characteristics in the FDR data is that there are the outliers in some parameters shown in Fig. 2 (a) and the measurement noises are widely corrupted in some sensor measurements shown in Fig. 2 (b).



(a) Aileron position (outlier) (b) Flight path acceleration (noise)

Figure 2. The examples of the parameters corrupted with the outliers and the measurement noises.

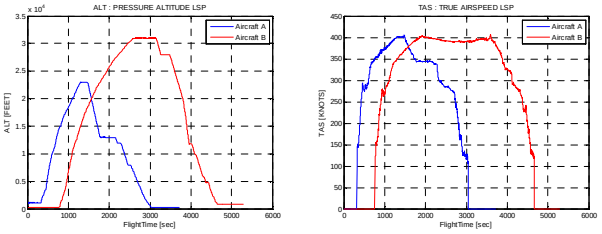


(a) Altitude (b) Body Longitudinal Acceleration

Figure 3. The examples of different value ranges of each parameter.

Another characteristic of the FDR data, each parameter has a considerably different value range because a physical quantity of each parameter is different shown in Fig. 3. The value range of altitude is about 0 to 25,000 and the value range of body longitudinal acceleration is about -0.2 to 0.3 in these examples. Additionally, sampling rates of each parameter are different.

Another specific characteristic of the FDR data used in this study is that flight patterns in the FDR data are different according to flight routes of aircraft. As an example, Fig. 4 (a) and (b) show the altitude and the airspeed of the FDR data regarding two different aircraft A and B. Here, we can readily observe that the aircraft A and the aircraft B have the considerably different flight patterns so that time series data between the FDRs data are not directly comparable.



(a) Altitude (b) Airspeed

Figure 4. The examples of selective parameters of different FDR data.

B. Data Pre-processing

1) Data Filtering

As mentioned above, the raw FDR data contain the outliers and the measurement noises. Thus, before getting into the data analytics process, these error sources should be filtered. In the presented development, two different filters (i.e., a moving median filter and a second-order linear filter) are sequentially applied as shown in Fig. 5. The purposes of these filters are given by

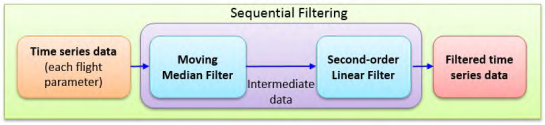


Figure 5. The diagram of data filtering process.

- **Moving median filter**: to suppress the outliers in the time series data
- **Second-order linear filter**: to reject the measurement noises in the time series data

Fig. 6 shows a process of moving median filter. As a window moves in the time series data, this filter takes a median value from values within the window. The performance of moving median filter is decided by the size of window. The best way to choose an appropriate windows size is based on a set of experiments since the performance of moving median filter highly depends on the properties of outliers in the time series data. In this study, the size of window is selected as $W = 13$ based on the set of experiments.

A second-order linear filter shown in Fig. 7 can be written as

$$\sum_{i=0}^{3} a_i y_{k-i} = \sum_{j=0}^{2} b_j x_{k-j} \quad (1)$$

where a_i is the i-th coefficients of denominator and b_j is the j-th coefficients of numerator, respectively. In this filter, these parameters are chosen to achieve a desired bandwidth of filter. Also, the best way to select an appropriate bandwidth of filter is based on a set of experiments. In this study, the desired bandwidth and damping ratio of second-order linear filter are selected as $\omega_n = 0.1\text{Hz}$ and $\zeta = 0.707$ in order to ensure a sufficient noise level reduction.

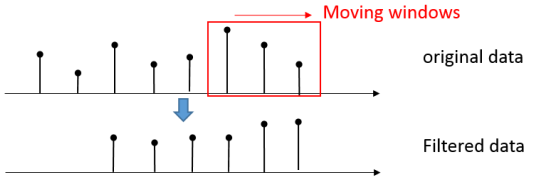


Figure 6. The moving median filter.

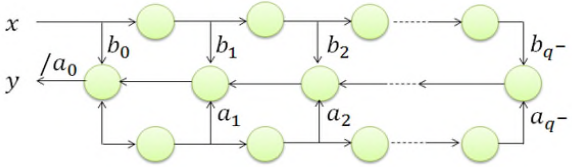
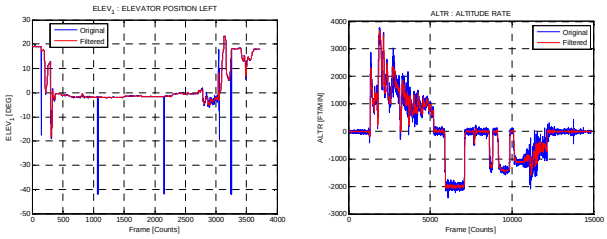


Figure 7. The second order linear filter.

Fig. 8 shows the illustrative example of the filtered time series data, where the blue line is the original time series data and the red line is the filtered time series data. As shown in Fig. 8, we can readily observe that outliers and measurement noises are successfully suppressed after passing two filters sequentially.



(a) Suppression of outliers (b) Rejection of measurement noises

Figure 8. The examples of cleaned data after passing the two filters.

2) Data Sampling

In the FDR data, each parameter is recorded throughout a flight. Accordingly, the flight patterns of aircraft are different because of their own flight routes. Additionally, each parameter is recorded with the different sampling rate. Accordingly, the time series data in the FDR are not directly comparable. Therefore, before getting into the feature generation, the FDR data should be mapped into a comparable data space, anchored by a specific event in time (in this paper, we call it as a region of interest, ROI). Additionally, it is also needed to match up the sampling rate of each parameter. To make the FDR data become comparable, the ROI should be chosen as common flight

phases. Namely, given the FDR data, each parameter should be sampled at a fixed time interval starting from the common anchoring event. In this paper, a final approach phase is considered as the common anchoring event because most of failures likely occur at this flight phase.

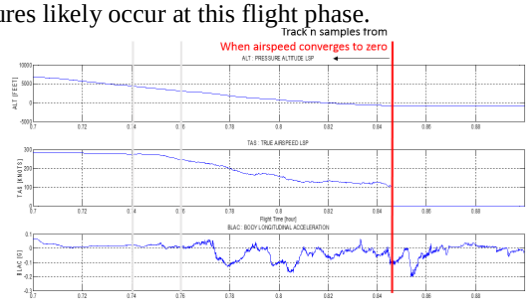
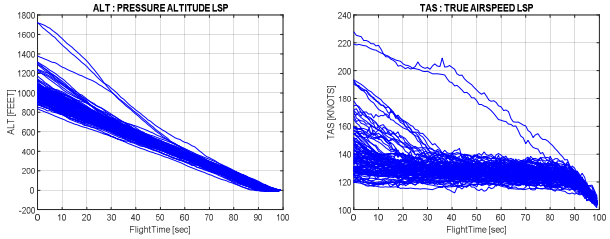


Figure 9. The sampling of time series data: the final approach phase.

During the final approach phase, the anchor is the point in time when the airspeed decreases to zero shown in Fig. 9. In this figure, the x-axis is the flight time in hour and the y-axis are the values of parameters: the altitude, the airspeed, and the axial acceleration, respectively. For each parameter, the sampled time series data is determined by tracking n samples from the anchor point backwardly. As an example, Fig. 10 shows the sampled time series data: the altitude and the airspeed.



(a) Altitude (b) Airspeed

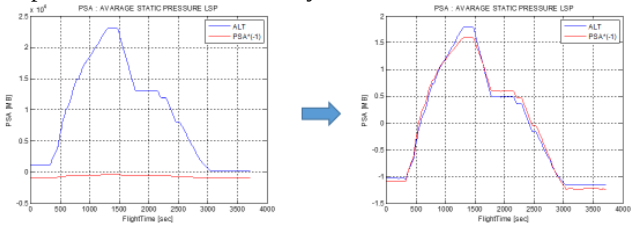
Figure 10. The examples of sampled time series data in the case of final approach phase.

3) Data Transformation

As mentioned before, the value ranges of each parameter are considerably different. If such parameters are used together, incorrect results are expected. Therefore, all the parameters are needed to be transformed in order to uniformalize their value ranges. To this end, in this paper, the $\mu-\sigma$ standardization method [6] is adopted. After the data sampling step, a $n\times 1$ time series vector for each parameter is determined. Each time series vector is then transformed as

$$x'_k = \frac{x_k - \bar{x}}{s_x}, \qquad \text{where} \quad k = 1,\ldots,n \tag{2}$$

where $\bar{x}$ represents the mean value of time series vector within the sampling interval and s_x is the variance. x_k and x'_k are the original data point and the transformed data point at time k, respectively. The additional effect of data transformation is that we can identify the parameters that have similar patterns. For example, Fig. 11 shows the time series data regarding the altitude and the negative value of pressure. The original time series data of two parameters look totally different shown in Fig. 11 (a). After the data transformation, however, the two parameters are matched shown in Fig. 11 (b). The reason behind the two similar patterns after the data transformation is that the pressure is negatively proportional to the altitude, physically. The data transformation can extract this physical pattern and it is helpful to the correlation analysis.



(a) Original time series (b) Transformed time series

Figure 11. The examples of data transformation.

C. Feature Generation

1) Correlation Analysis

In this development, the feature is chosen as the high-dimensional vector formed by arranging all the sampled parameters. A comparison of feature vectors then detects the FDR data that contain the unusual patterns. Therefore, in this concept, the effect of multi-collinearity between the parameters can reduce the analysis performance. Therefore, in order to determine the redundant information in the continuous parameters, the correlation analysis is performed in advance, using the Pearson correlation coefficient [6]. Suppose there are two $n\times 1$ time series data as

$$\begin{aligned} \underline{x}^i &= \left[x_1^i, x_2^i, \ldots, x_k^i, \ldots x_n^i \right] \\ \underline{x}^j &= \left[x_1^j, x_2^j, \ldots, x_k^j, \ldots x_n^j \right] \end{aligned} \tag{3}$$

Then, the Pearson correlation coefficients for the two time series $S_{i,j}$ are given by

$$S_{i,j} = \frac{\sum_{k=1}^{n} \left(x_k^i - \bar{x}^i \right)\left(x_k^j - \bar{x}^j \right)}{\sqrt{\left(\sum_{k=1}^{n} \left(x_k^i - \bar{x}^i \right)^2 \right)\left(\sum_{k=1}^{n} \left(x_k^j - \bar{x}^j \right)^2 \right)}} \tag{4}$$

where $\bar{x}^i$ represents the mean value of the i-th time series data and $\bar{x}^j$ represents the mean value of the j-th time series data. Dually, x_k^i and x_k^j are the values at time k. Fig. 12 shows a heatmap of the Pearson correlation coefficients between all the continuous parameters, where the x-axis and the y-axis are the index of each parameter. As shown in this figure, we can readily observe that some continuous parameters are linearly correlated.

Through the correlation analysis, several groups of correlated parameters are identified. As an illustrative example, an one group of correlated parameters is provided in Table I. Fig. 13 shows the patterns of those parameters and confirms that the patterns of those parameters are highly correlated. The reason behind is that the physical quantities of those parameters are implicitly related with the altitude. Therefore, after the data transformation, we can reveal the correlations of those parameters.

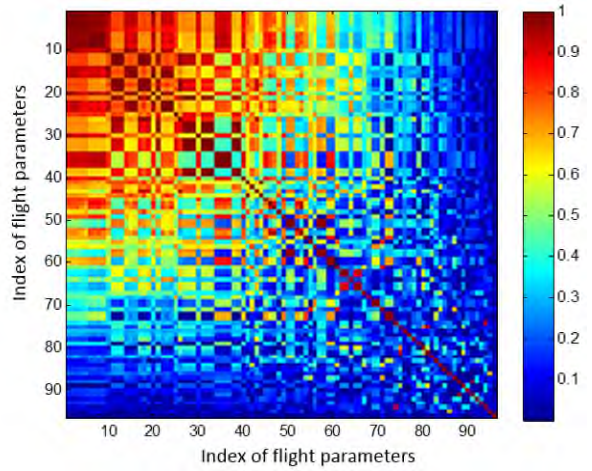


Figure 12. The heatmap of Pearson correlation coefficient between all the continuous parameters.

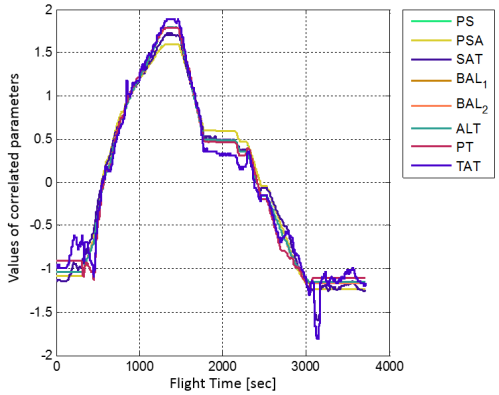


Figure 13. The time series data of correlated parameters.

TABLE I. THE DESCRIPTION OF CORRELATED PARAMETERS

Parameters	***Descriptions***
PS	Static Pressure
PSA	Average Static Pressure
SAT	Static Air Temperature
BAL1	Baro Correct Altitude 1
BAL2	Baro Correct Altitude 2
ALT	Pressure Altitude
PT	Total Pressure
TAT	Total Air Temperature

2) Correlation Relaxation

As observed in the previous section, some parameters are highly correlated. Therefore, a correlation relaxation process is required to weaken the effect of multi-collinearity. In order to relieve the correlations, as in a similar way provided in [4], the correlated parameters are compressed into the mean values among the correlated parameters. Suppose there are numbers of $n \times 1$ time series data which are highly correlated each other, as

$$G = \left\{ \underline{x} \in \mathbb{R}^{n \times 1} \middle| \underline{x}^1, \underline{x}^2, \ldots, \underline{x}^{n_c} \right\} \tag{5}$$

where n_c is the total number of correlated parameters. Then, at time k, the mean value of correlated parameters is given as

$$\overline{x}_k = \frac{1}{n_c} \sum_{i=1}^{n_c} x_k^i \tag{6}$$

where $\overline{x}_k$ represents the mean value and x_k^i represents the value of the i-th correlated parameter at time k. After taking the correlation relaxation process for all the groups of correlated parameters, the Pearson correlation coefficients for all the continuous parameters are determined again. Fig. 14 shows the heatmap of the Pearson correlation coefficients after the correlation relaxation, where the x-axis and the y-axis are the index of each parameter. As shown in Fig. 14, we can readily observe that the effect of multi-collinearity is greatly weakened. The additional effect of this process is to reduce a number of parameters so that the dimensionality of feature vectors can be reduced.

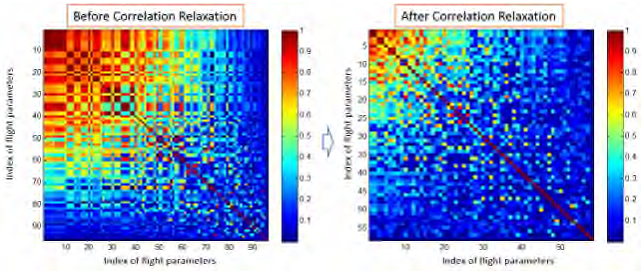


Figure 14. The heatmap of the Pearson correlation coefficients between all the continuous parameters (left: before, right: after) .

3) Feature Vector Generation

For each FDR data, the sampled time series data are arranged to form three feature vectors according to the types of parameters: the continuous signal, the discrete signal, and the warning signal, respectively. In this analysis, we consider the continuous parameters and the discrete parameters separately since the characteristics of these parameters are significantly different. Namely, when we compare the feature vectors, different similarity measures are required for each parameter type. Also, the reason why the warning parameters are separately handled is that when one of warning occurs, that FDR data should be carefully investigated even though other parameters seem to be okay. In the FDR data, there are a large number of the continuous parameters. Compared to the continuous parameters, there are few of the discrete parameters and the warning parameters.

First, the high-dimensional vector formed by the continuous parameters is written as

$$\underline{c}_k = \left[x_1^1, x_2^1, \ldots, x_n^1, \ldots, x_j^i, \ldots, x_n^{m_c} \right] \tag{7}$$

where $\underline{c}_k$ is the feature vector formed by the continuous parameters for the k-th FDR data. m_c is the number of the continuous parameters. In Eq. (7), x_j^i represents the value of the i-th continuous parameter at sample time j. Therefore, the total dimensionality of this vector is given $m_c \times n$ and it is usually high. For all the FDR data, the feature vectors for the continuous parameters are arranged to form a dataset as

$$D_c = \begin{bmatrix} \underline{c}_1 \\ \underline{c}_2 \\ \vdots \\ \underline{c}_{N_t} \end{bmatrix} \quad (8)$$

where N_t is the number of FDR data used in this analysis. In a similar way, the feature vectors for the discrete parameters $\underline{d}_k$ and the warning parameters $\underline{s}_k$ can be determined. Dually, for all the FDR data, the dataset for the discrete parameters D_d and the warning parameters D_s can be computed.

4) Dimension Reduction

The dimensionality of feature vector for the continuous parameters is considerably high. Also, the feature vectors for all the FDR data are sparsely distributed across the dimensions. Such the high-dimensionality and the data sparseness problem lead to incorrect results. In order to relieve these problems, we use the PCA [6] in order to transform the original dataset into a new orthogonal coordinate system that maximizes their variances. In order to reduce the dimension, we keep the first few components with the majority of the information in the new coordinate system. Accordingly, we select down the first largest K principal components that capture 90% of the variance in the dataset, as

$$\frac{\sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{m_c} \lambda_i} > 0.9 \quad (9)$$

where λ_i is the variance in the new coordinate system, which is computed by using the dataset for the continuous parameters in conjunction with the singular value decomposition (SVD) method. After performing the PCA, a new dataset for the continuous parameters in the new coordinate system is obtained as

$$D_c' = \begin{bmatrix} \underline{c}'_1 \\ \underline{c}'_2 \\ \vdots \\ \underline{c}'_{N_t} \end{bmatrix} \quad (10)$$

with

$$\underline{c}'_{(\cdot)} = \left[x'_1, x'_2, \ldots, x'_i, \ldots, x'_K \right] \quad (11)$$

where x'_i represents the i-th component of vector in the new coordinate system.

D. Detection Algorithm for Unusual Patterns

In this approach, the unusual patterns are identified by examining the distance to a point's k-nearest neighbors. Fig. 15 illustrates the process of k-nearest neighbor approach in the case of $k = 5$ [6]. As shown in Fig. 15, if a point's neighboring points are relatively close, then that point (blue) is considered as the normal one. Otherwise, if a point's neighboring points are not adjacent, then, that point (red) is considered as the outlier. In this approach, the proximity the between feature vectors for each FDR data is determined by using a distance measure. In this study, the weighted sum of distance measure for each type of parameters is considered, as

$$L(i,j) = w_c \phi_c \left(\underline{c}'_i, \underline{c}'_j \right) + w_d \phi_d \left(\underline{d}_i, \underline{d}_j \right) + w_s \phi_s \left(\underline{s}_i, \underline{s}_j \right) \quad (12)$$

where $L(i,j)$ denotes the distance between the i-th FDR data's feature vector and the j-th FDR data's feature vector. In this study, we take the kernel matrix as the weighted sum of similarity measure of the continuous, the discrete and the warning parameters. $\phi_c(\bullet)$, $\phi_d(\bullet)$, and $\phi_s(\bullet)$ represent the similarity measures for each type of parameters. w_c, w_d, and w_s denote the weights of each type of parameters. In this study, a large weight value is assigned to the warning parameters as $w_s \gg w_c, w_d$ because those parameters are important and critical indicators for the anomaly detection. For the continuous parameters, the Euclidean distance is chosen as the similarity measure. For the discrete parameters, the longest common subsequence (LCS) [8] is considered as the similarity measure. For the warning parameters, the Jaccard distance [9] is used.

Generally, the k-nearest neighbor approach leads to a high computational complexity since the distances with each data point should be determined. In order to reduce the computational complexity, in this study, we apply the k-nearest neighbor algorithm in conjunction with randomization and simple pruning rule, called the ORCA [10].

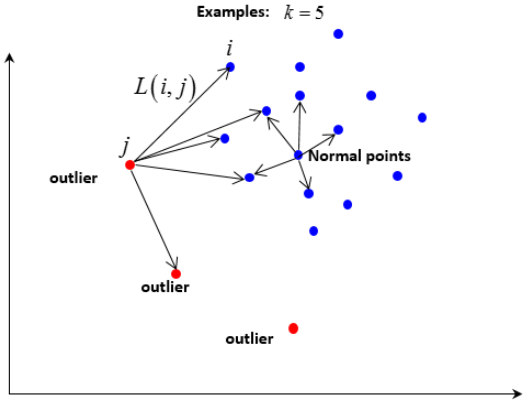


Figure 15. The illustration of k-nearest neighbor approach [6].

IV. Data Analysis Results

In this section, the proposed data analytics is carried out using the realistic FDR data provided from the NASA's open data base [7]. In this paper, 2,169 numbers of FDR data are utilized. Then, the detection of FDR data that contain the unusual patterns is performed using the ORCA [10]. The design parameter of ORCA is chosen as $k = 10$. Also, the desired detection rate is chosen as 2%.

Fig. 16 shows the detection results of ORCA, where the x-axis represents the index and the y-axis represents the similarity distance of ORCA algorithm. Fig. 17 provides the scatter plot of feature vectors in the principal coordinate system. In this figure, the first 3 principal components are

only selected to map the high-dimensional feature vectors into the three-dimensional space. In this figure, circle points represent the abnormal feature vectors and “x”-shaped points are the normal feature vectors. The results shown in Figs. 16 and 17 indicate that the proposed method can successfully identify the standalone feature vectors. Through this analysis, 40 FDR data that contain the unusual patterns are identified.

The proposed algorithm also reports that there might be the unusual patterns in some parameters such as the altitude, the airspeed, the airbrake position, and so on. In order to confirm the detection results, we draw those parameters as shown in Fig. 18. As shown in Fig. 18 (a) and (b), the unusual patterns of altitude and airspeed are slightly higher than the normal patterns. As shown in Fig. 18 (c) and (d), the negative values in ABRK and AT, which are unusual, are identified. In Fig. 18 (e) and (f), the delay of landing gear and the flap are detected. Also, the abnormal warning signals are automatically identified as shown in Fig. 18 (g) and (h). Accordingly, the results indicate that the proposed algorithm can successfully detect the unusual patterns in the FDR data, regardless of the parameter types: the continuous, the discrete, and the warning signals. Therefore, the proposed method can help a human operator to consume all the FDR data within a limited time. It can be utilized for a high-level fault diagnosis for enhancing the airline maintenance operations.

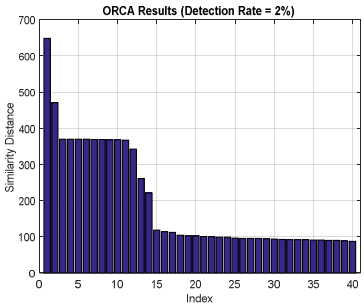


Figure 16. The Similarity distances of detected feature vectors.

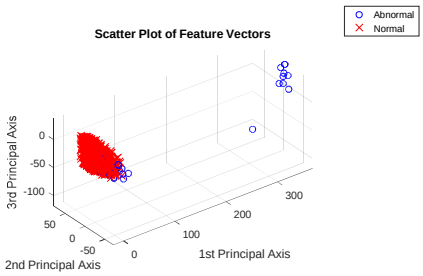
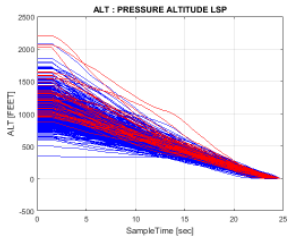
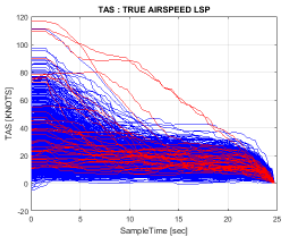


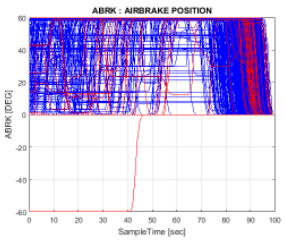
Figure 17. The scatter plot of feature vectors.



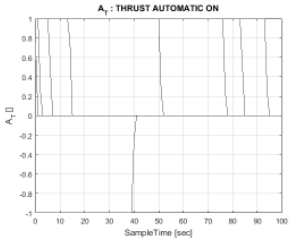
(a) Altitude



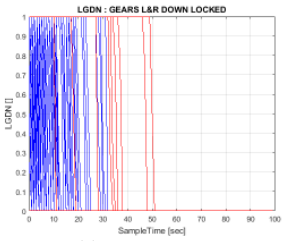
(b) Airspeed



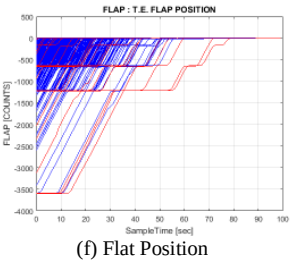
(c) Airbrake Position



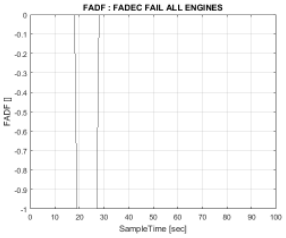
(d) Thrust Automatic On



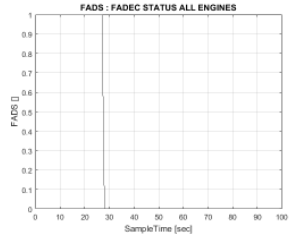
(e) Gears Down



(f) Flat Position



(g) Fail All Engines



(h) FADEC Status All Engines

Figure 18. The time series data that have the unusual patterns**.**

V. Conclusion

In this paper, the data analytics development to identify the unusual pattern of flights from a large amount of the FDR (flight data recorder) data is proposed for improving the airline maintenance operations. To this end, in the data pre-processing step, the data filtering, the data sampling, and the data transformation are conducted. The three types of parameters, such the continuous signal, the discrete signal, and the warning signal, are handled separately. The high-dimensional vector formed by arranging the time series data is considered as the feature. The correlation analysis, the correlation relaxation, and the dimension reduction are sequentially performed to generate the feature vector. For each type of parameters, the three types of feature vectors are generated. And, a type of k-nearest neighbor algorithm is used with the weighted sum of similarity distances for each type of parameters. The Euclidian distance, the longest common subsequence, and the Jaccard distance are utilized for the similarity measures of each type of parameters. The proposed method is tested with using the realistic FDR data from the NASA’s open database. The numerical analysis results indicate that the proposed approach can be used to automatically identify the FDR data in which the unusual patterns are recorded from a large amount of the FDR data. Therefore, the proposed method can be used for a high-level fault diagnosis in the airline maintenance operations.

REFERENCES

[1] E. A. Stephenson, "Aircraft flight data recorder data acquisition system," *US Patent 4656585*, 1987.

[2] S. Budalakoti, A. N. Srivastava, and R. Akella, "Discovering atypical flights in sequences of discrete flight parameters," *IEEE Aerospace Conference*, 2006, pp. 1-8.

[3] S. Das, B. L. Matthews, A. N. Srivastava, and N. C. Oza, "Multiple kernel learning for heterogeneous anomaly detection: algorithm and aviation safety case study," *Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining*, 2010, pp. 47-56.

[4] L. Li, S. Das, R. John Hansman, R. Palacios, and A. N. Srivastava, "Analysis of flight data using clustering techniques for detecting abnormal operations," *Journal of Aerospace Information Systems*, vol. 12, no. 9, 2015, pp. 587-598.

[5] M. Ester, H.-P. Kriegel, J. Sander, X. Xu et al., "A density-based algorithm for discovering clusters in large spatial databases with noise," *Knowledge Discovery and Data mining*, vol. 96, no. 34, 1996, pp. 226-231.

[6] T. A. Runkler, *Data Analysis*, Springer , 2012.

[7] B. Matthews, "Flight Data for Tail 653," NASA DASH link (https://c3.nasa.gov/dashlink/resources/631)

[8] J. W. Hunt and T. G. Szymanski, "A Fast Algorithms for Computing Longest Common Subsequence," *Communications of the ACM*, vol. 20, no. 5, 1977, pp. 350-353.

[9] P. Jaccard, "The distribution of the flora in the alpine zone," New *phytologist*, vol. 11, no. 2, 1912, pp. 37-50.

[10] S. Bay and M. Schwabacher, "Mining distance-based outliers in near linear time with randomization and a simple pruning rule," *Proceedings of SIGKDD*, 2003, pp. 29-38.

2017-12-11

Data analytics development of FDR (Flight Data Recorder) data for airline maintenance operations

Lee, Chang-Hun

IEEE

Chang-Hun Lee, Hyo-Sang Shin, Antonios Tsourdos and Zakwan Skaf. Data analytics development of FDR (Flight Data Recorder) data for airline maintenance operations. Proceedings of the 2017 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI 2017), 16-18 November 2017, Daegu, South Korea.

http://dx.doi.org/10.1109/MFI.2017.8170443